DOCUMENT RESUME

ED 345 520 FL 020 215

AUTHOR	Tierney, Joseph; Mack, Molly
TITLE	Vocoders and Speech Perception: Uses of Computer-Based Speech Analysis-Synthesis in Stimulus Generation.
PUB DATE	87
NOTE	19p.; For the journal as a whole, see FL 020 212.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Journal Articles (080)
JOURNAL CIT	IDEAL; v2 spec iss p45-61 1987
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Applied Linguistics; *Auditory Perception; *Auditory Stimuli; *Computer Oriented Programs; *Computers; Language Processing; *Language Research; *Speech; Speech Synthesizers
IDENTIFIERS	*Vocoders

ABSTRACT

Stimuli used in research on the perception of the speech signal have often been obtained from simple filtering and distortion of the speech waveform, sometimes accompanied by noise. However, for more complex stimulus generation, the parameters of speech can be manipulated, after analysis and before synthesis, using various types of algorithms to generate stimuli of interest to the researcher. There are several computer-based analysis and synthesis techniques, or vocoders, that make different assumptions about the speech production and reception mechanisms. These include: (1) the channel vocoder, which uses filter banks to split the speech spectrum into discrete frequency ranges measured often enough to keep up with speaker's articulation rate; (2) the homomorphic vocoder, which derives source and vocal tract parameters from the spectrum signal alone; (3) the linear predictive coding vocoder, which generates the speech signal in a manner similar to the two previous vocoders, but measures the vocal tract by fitting the speech waveform in some interval to the output waveform of a linear filter; and (4) the sinusoidal transform coding vocoder, which, rather than treating the sound source as a periodic pulse train or noise signal, characterizes it as a sum of sinusoidal waves that vary with respect to one another. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED345520

and Speech Perception:
Vocoders omputer-Based Speech Analysis-Synthesis
Uses of Cus Generation
in Stimulu

Joseph Tierney and Molly Mack

OF EDUCATION
U.S. DEPARTMENT arch and Improvement
Office of Educational Rese CES INFORMATION
EDUCATIONAL RESOUR (ERIC)
CENTER (een reproduced as
son or organization
★ This document has b
received from the perso made to improve
originating it
☐ Minor changes have be
reproduction quality
s stated in this docu-
ly represent official
• Points of view or opinion
ment do not necessaril
OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Wayne Dickerson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL020 915

BEST COPY AVAILABLE

IDEAL 2, 1987

VOCODERS AND SPEECH PERCEPTION: USES OF COMPUTER-BASED SPEECH ANALYSIS-SYNTHESIS IN STIMULUS GENERATION

Joseph Tierney and Molly Mack

Work on the perception of the speech signal has produced much useful psychophysical data. In previous years, stimuli used to produce these data have been obtained from simple filtering and distortion of the speech waveform, sometimes accompanied by noise. For purposes of more complex stimulus generation however, the parameters of speech can be manipulated--after analysis and before synthesis--using various types of algorithms in order to generate stimuli of interest to the experimenter.

In this article, we discuss several computer-based analysis-synthesis techniques which make different assumptions about the speech production and perception mechanisms. We also describe experiments in which synthesized speech has been or easily could be utilized.

INTRODUCTION

Experiments in psychophysics and speech perception which use speech or speech-like signals as stimuli are not always easy to design and administer because of the variability of the signal and the difficulty of isolating and controlling single features. In spite of the challenges posed by using speech stimuli, experiments utilizing natural speech which has been filtered and distorted have produced valuable data over the past several decades (Miller & Nicely, 1954; Pisoni & Hunnicutt, 1980; Luce et al., 1983; Pisoni et al., 1985).

Coexistent with this activity in auditory phonetics and speech generation has been work in speech coding for communications applications. The computer-based techniques used in this work can be adapted readily to generate interesting speech stimuli for perception research. More specifically, transforming the acoustic signal into a descriptive stream of binary digits suitable for transmission over telephone lines, satellite links, and packet communications has been a major activity of communications engineers for many years (Flanagan et al., 1979). Many techniques have been developed for dealing with this type of speech coding. All of these techniques can be implemented with programs running on current or proposed digital processors, and therefore they can be used in speech-perception and psychophysical research.

SPEECH ANALYSIS-SYNTHESIS AND APPLICATIONS

Speech encoding devices may be classified into three categories according to their data rate--i.e., the rate at which they transmit a binary signal. High data-rate devices, such as those in the 16-to-64 kilobits/sec range, are not very complicated. They work directly on the speech waveform and they utilize very little knowledge

about the underlying speech mechanisms. These devices are used as standards in the telephone industry and in the military. Mid data-rate devices, such as those in the 4.8-to-16 kilobits/sec range, are more sophisticated than the high data-rate devices, and they have some potentially interesting applications in speech-perception research. It is the low data-rate devices, however, which are perhaps of greatest interest from the speech researcher's point of view. These devices, often referred to as vocoders or analysis-synthesis devices, have data rates in the range of .6 to 9.6 kilobits/sec. Vocoders provide the greatest potential for speech signal manipulation because they assume a speech-production model in analysis-synthesis.

For speech to be usefully intelligible, as in standard telephone communication, it must contain acoustic information up to about 3.5 kilohertz (kHz).[1] (As a point of reference, a human with good hearing can perceive frequencies of approximately 20 Hz to 20 kHz--i.e., from 20 to 20,000 cyles/sec.) In order to encode and transmit a speech signal with a bandwidth of 4-5 kHz at 2000-3000 bits/sec, it becomes necessary to model the speech production mechanism and to measure parameters of the model. This process is called analysis. At the receiving end of the transmission, the measured parameters are used with the model to recreate or synthesize the speech signal. This process of analysis-synthesis is the basis for several voice-coding algorithms which allow a speech signal to be encoded and transmitted in the range of 1000-4000 bits/sec.

The analysis process presents us with parameters which describe the speech signal and which can be modified to alter certain of its properties. Almost all of the systems in use at present, and many of those proposed for future use, assume that the speech-production mechanism can be modeled as in Figure 1. This mechanism has a sound source and a filter. The sound source may be the vocal cords which produce voiced (periodic) sounds and which determine the fundamental

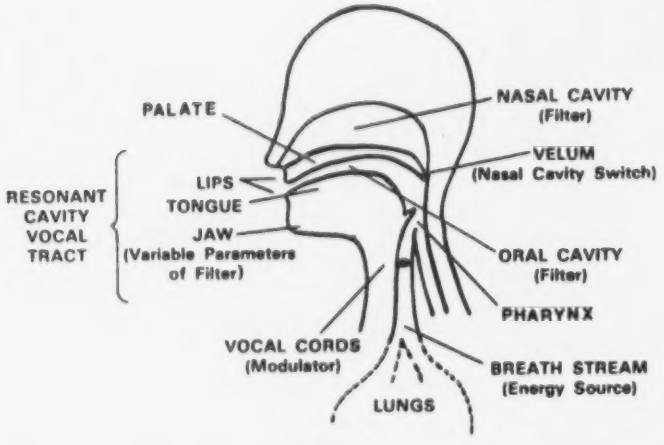


Figure 1. Lateral view of the speech-production mechanism.

frequency--recognized perceptually as pitch. The source may also be some type of intraoral constriction, such as that used to produce voiceless (aperiodic) sounds, or it may be a combination of voiced and unvoiced sounds, such as that used to produce a voiced fricative like [v]. The filter is the supralaryngeal vocal tract which includes the nasal, oral, and pharyngeal cavities. It may properly be considered a variable acoustic filter whose shape specifies the formant frequencies (also called the vocal tract resonances) of a sound. In Figure 2, note the display of a simple input waveform (a waveform generated by the energy or sound source) on the left. The period of the waveform is represented by T. To the right is the complex output waveform as it appears after being filtered by the supralaryngeal vocal tract (resonant cavity). Below these waveforms are spectra displaying the fundamental frequency (F0) and harmonics of the sound source, the spectrum envelope with formant frequency peaks, and the actual output signal, with the spectrum envelope superimposed upon the source signal.

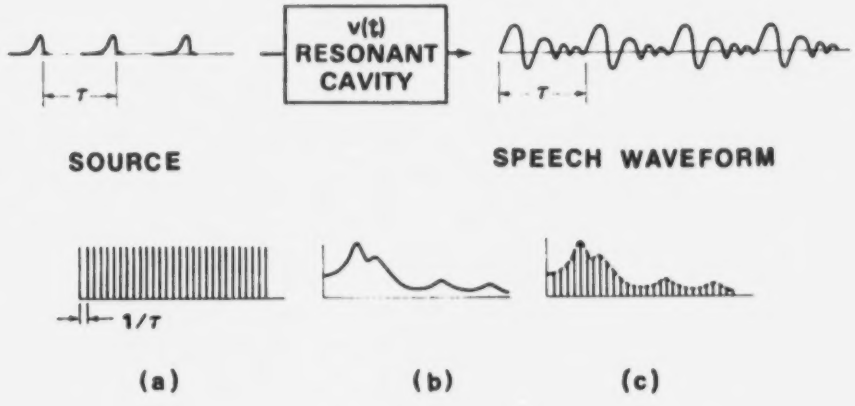


Figure 2. Top line: Input signal before and after filtering by supralaryngeal vocal tract (resonant cavity). Bottom line: Spectra showing fundamental frequency and harmonics generated by sound source (a); spectrum envelope with filter formant frequency peaks (b); and combined source and filter spectrum (c).

The challenge in speech analysis is to measure accurately these speech attributes using only the transduced acoustic signal available as the electrical output of a microphone. Assuming that the production model is accurate and that the articulators vary as a function of time, we can expect the measured spectra of speech signals to resemble the examples presented in Figure 3. The spacing of the harmonics (the vertical lines) in each spectrum reflects the properties of the sound source, with more widely spaced lines characterizing a signal of higher frequency, as in (a), and more narrowly spaced lines characterizing a signal of lower frequency, as in (b). The vocal tract filter determines the envelope of the

spectrum whose peaks, at or near formant frequencies, specify a given sound. The height of the peaks reflects their relative amplitude, usually measured in decibels. In this figure the spectrum envelope is of a different shape in (b) and (c) because the configuration of the vocal tract used to produce the sounds in (b) and (c) was different. Thus, the spectrum in (a) could be the vowel [a] produced by an adult female, and the spectrum in (b) could be that of the same vowel produced by an adult male. The spectrum in (c) could be that of the vowel [i], also produced by an adult male. Once the relationship between the source-filter mechanism and the spectrum is understood, we can estimate the acoustic properties of the source and vocal tract filter from the measured spectrum. In fact, this is one way of deriving measurements for vocoded speech.

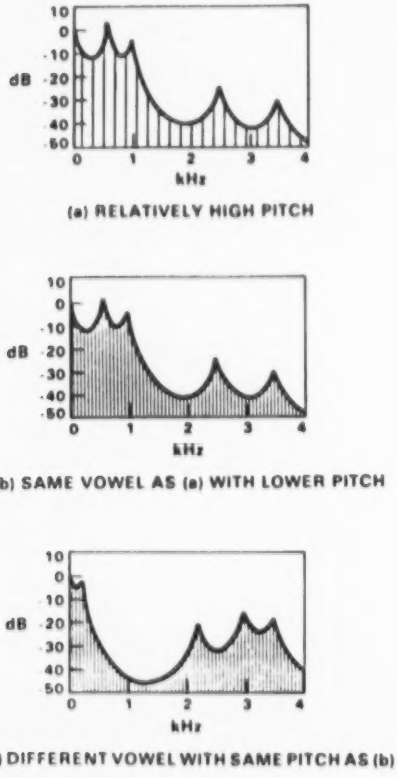


Figure 3. Line spectra showing fundamental frequency, harmonics, and formant frequencies for higher- and lower-pitched vowels.

In the sections which follow, we shall discuss several types of vocoders, and we shall consider specific experiments which have been, or could be, conducted using these vocoders. Although the channel vocoder will receive the greatest amount of attention, we do not mean to suggest that it is better suited to speech research than are the other types of vocoders. In reality, a number of the experiments discussed below could be carried out with any of the vocoders under consideration. (In fact, the selection of one type of vocoder over another may simply reflect the personal preference of the researcher.) However, to the extent that each vocoder utilizes different algorithms for analysis and synthesis, each will produce speech sounds of somewhat different quality. And, more importantly, each vocoder will control the synthesized output in a specific way, permitting different types of manipulation. Thus, it may be that one type of vocoder is somewhat more effective for the exploration of certain research issues (e.g., perturbation of the spectrum or speech produced in noise) than another.

Channel Vocoder

One typical analysis-synthesis speech coder is the channel vocoder, presented schematically in Figure 4. In the channel vocoder, an analyzer consisting of a bank of filters splits the speech spectrum into discrete frequency ranges whose energies are measured often enough to keep up with the speaker's rate of articulation (Tierney et al., 1964; Gold et al., 1983). When the signal is synthesized, its spectrum is recreated using a similar filter bank. In the measurement of the signal needed to recreate the sound source of the original signal, the waveform can be examined directly, or the signal spectrum can be analyzed. If the waveform is used, only a low-frequency portion is needed to determine if the source is the vocal cords (as in a voiced sound, such as [d] or [a]), if it is a noise-producing constriction (as in a voiceless sound, such as [t] or [s]), or if it is an unconstricted egressive air flow (as in [h]).

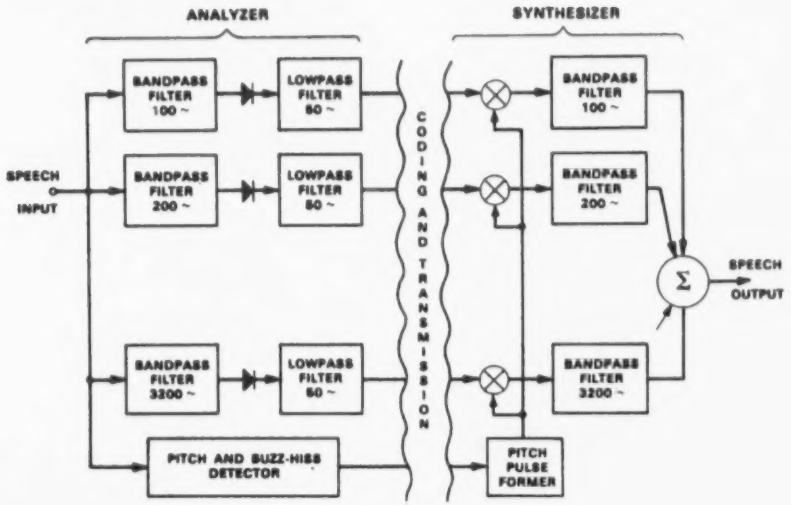


Figure 4. Block diagram of a typical channel vocoder.

With the spectrum envelope (whose frequencies approximate those of the vocal tract) and the spectrum fine structure (whose frequencies approximate those of the sound source), it is possible to recreate a speech signal by measuring the pertinent parameters rapidly (every 10-15 msec). What we cannot retain and recreate with this model is the detailed timing relationship between the different frequency components; that is, we measure only the energy in bands, not the relationship between the harmonics which constitute the signal. However, if the filter bank has sufficient resolution--i.e., if the frequency range of its filters is narrow enough--and if the source determination is sufficiently accurate, we can still obtain recreated speech of excellent quality.

The channel vocoder has a number of interesting applications in speech research, and several experiments have been conducted recently which illustrate some of these applications. For example, the sound source itself is of interest. The pitch signal used in the synthesizer contains very little information about the vocal tract, but it characterizes the fundamental frequency (F0) of the speaker's voice. The waveform of this signal can be used as a basis for pitch discrimination experiments in which the fundamental frequency is altered in various ways. A simple distortion can, for example, be obtained by adding or multiplying the F0 by a constant. The resulting pitch signal can then be passed through the vocal tract filter to determine what the speech sounds like with this distortion. One obvious application of this procedure involves the separation of source and filter. A researcher might wish to examine the perceptual effect of a high-pitched (e.g., female) voice being produced by a large (e.g., male) vocal tract. In doing so, one could thus ascertain the relative perceptual salience of fundamental and formant frequencies in signaling the gender of a speaker and one could determine what kind of trading relation exists between these two acoustic cues.

Another application involving manipulation of the sound source may be found in experiments in which listeners' ability to discriminate pitch is observed. Such an experiment has been conducted by Mack and Gold (1984; 1986). In this experiment, listeners were presented with three different types of stimuli consisting of simple pulse trains (non-speech stimuli with some acoustic complexity), monotone sentences, and sentences with natural intonation. For the generation of the sentence stimuli, Mack and Gold used a high-quality channel vocoder to produce sentences whose F0 differed from 0 to 6 Hz in .5 Hz increments. Subjects were presented with pairs of stimuli and were required to indicate whether the two members of each pair had identical or different pitch. Results are presented in Figure 5.

As is apparent from the figure, pitch discrimination, as determined by the percent of stimulus pairs judged as different, proved most accurate in pulse trains, and least accurate in naturally intoned sentences. These results suggest that the presence of linguistic information results in a relative decrement in listeners' sensitivity to fine differences in overall F0. Moreover, when a speech stimulus has a constantly changing F0 (as natural speech does), overall differences in the pitch of sentence pairs are especially difficult to detect. For, as may be seen in the figure, subjects performed near chance (with only about 50% of the pairs judged as different) in the natural-intonation condition, even in response to those stimulus pairs whose pitch was most dissimilar.

Mack and Gold used the same vocoder in a pilot experiment to determine listeners' responses to another type of permutation in the sound source. In this experiment, they generated pairs of sentences in which one member of the pair had

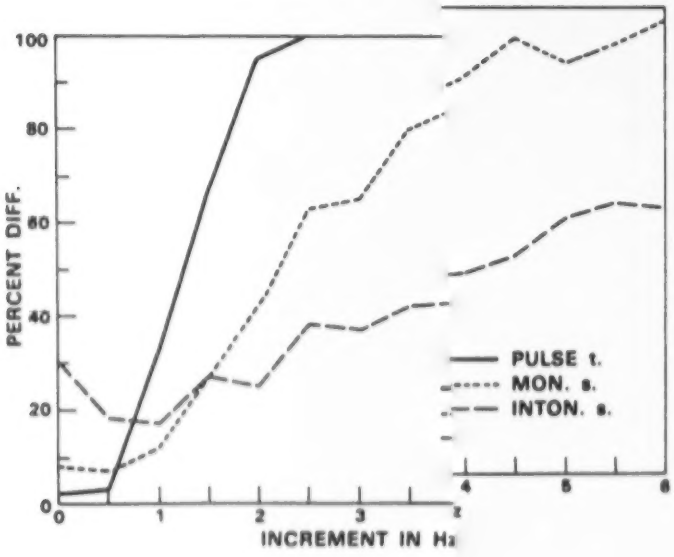


Figure 5. Discrimination functions for three types of stimuli--monotone pulse trains (PULSE t.), monotone sentences (MON. s.), and naturally intoned sentences (INTON. s.).

a natural intonation contour and the other had a contour which was linearly interpolated in increments of 20 to 140 msec. In Figure 6a, a schematic of this interpolation process is presented. As this figure illustrates, the F0 of the signal was linearly interpolated between two points, rather than being allowed to fluctuate in a natural manner. In a discrimination task using these stimuli, subjects were required to indicate whether they perceived a difference between the members of a stimulus pair in which one member had a natural intonation contour and the other an interpolated one. The objective of this study was to determine whether or not subjects would find sentences with interpolated pitch segments acceptable. Results of this experiment were, unfortunately, somewhat unclear due to the fact that some of the listeners found it extremely difficult to discriminate between the stimuli. Still, the trend was that subjects perceived a difference between the two types of stimuli, but only when longer interpolations were used (see Figure 6b). Thus, this experiment suggested that listeners seem to be relatively insensitive to changes in F0 provided that the overall pitch contour of a sentence is maintained--a finding in agreement with that of Collier (1975), t'Hart (1981) and de Pijper (1983).

BEST COPY AVAILABLE

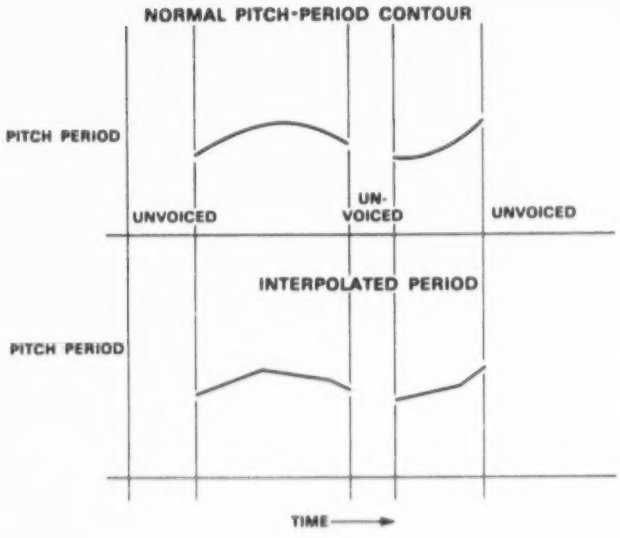


Figure 6a. Schematic of normal pitch-period contour and linearly interpolated pitch-period contour.

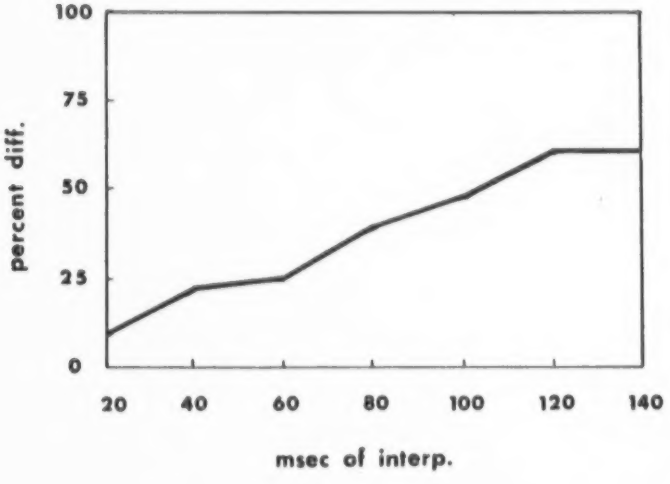


Figure 6b. Average discrimination function for pitch-interpolation experiment.

An additional experimental application of vocoders involves examination of the actual effect of vocoding itself. In a recent study (Mack and Gold, 1985), the intelligibility of speech in four conditions was analyzed. The conditions included (1) natural (non-vocoded) speech; (2) vocoded speech; (3) vocoded speech with noise added to the pitch track; and (4) vocoded speech with noise added to the spectrum. Speech stimuli consisted of 57 sentences presented auditorially to English monolingual subjects. All sentences were semantically anomalous--i.e., grammatical but meaningless. (An example of such a sentence is, "Simple rocks fear crooked hats.") Subjects were required to write each sentence as accurately as possible immediately after they had heard it. Analysis of the errors revealed that subjects made over twice as many errors in the vocoded as in the natural-speech condition, in spite of the fact that the vocoded speech was of extremely high quality. Moreover, they made numerous errors in response to vocoded speech with noise added to the pitch track, and they made even more errors when noise was added to the spectrum. Mack (1987) also used natural and vocoded anomalous sentences in a comparative study of English monolinguals and German-dominant German-English bilinguals. Analysis of the two groups' performance revealed that both found the vocoded sentences less intelligible than the natural ones. However, the bilinguals had an especially high error rate (even in response to the natural sentences, on which they made an average of six times as many errors as the monolinguals did). These findings reveal, in some detail, the effect of vocoded speech on perception--effects which might not be apparent in tests whose perceptual demands are slight.

Due to the fact that the analysis-synthesis vocoding procedure permits researchers to extract and manipulate source and filter information independently, there is another type of signal permutation--time expansion and compression--which has considerable potential in perception research. Specifically, an interesting consequence of time expansion and compression is that the temporal quality of an utterance can be changed without altering its fundamental or formant frequencies. To appreciate the significance of this, it is only necessary to consider what usually happens when a speech signal is played on a tape recorder at a faster or slower rate than that at which it was recorded. The speed of all speech events is changed and, as a result, the fundamental frequency of the speech is changed as well, resulting in an unnatural and high-pitched "chirpy" quality in the rapid speech and in a low-pitched "drunken" quality in the slow speech. While vocoding does permit one to make often dramatic speech rate changes in the output, the unnatural effects just described do not obtain, as is demonstrated in Figure 7. That is, the speed of the speech articulators is effectively decreased, but the frequencies are not changed. The resulting speech remains highly recognizable and the F0 is unperturbed. This procedure has considerable research potential and can be used in studies in which speech rate must be varied independently of frequency changes (which may, at the perceptual level, interact with subjective judgments about rate).

Homomorphic Vocoder

Another type of vocoder is the homomorphic vocoder. This vocoder derives source and vocal tract parameters from the spectrum signal alone (see Appendices 1a and 1b[2]). Assuming that the sound source and vocal tract spectra do not overlap badly, the logarithm of the spectrum can be used to turn a product into a sum. That is, if the parameters of the vocal tract (the variable acoustic filter) do not change rapidly relative to the sound source, then the vocal tract filter can be separated from the overall speech signal (Oppenheim, 1969). The synthesis procedure for this kind of analysis-synthesis is somewhat different from that of the channel vocoder. A filter bank is not used for recreating the vocal tract filter.

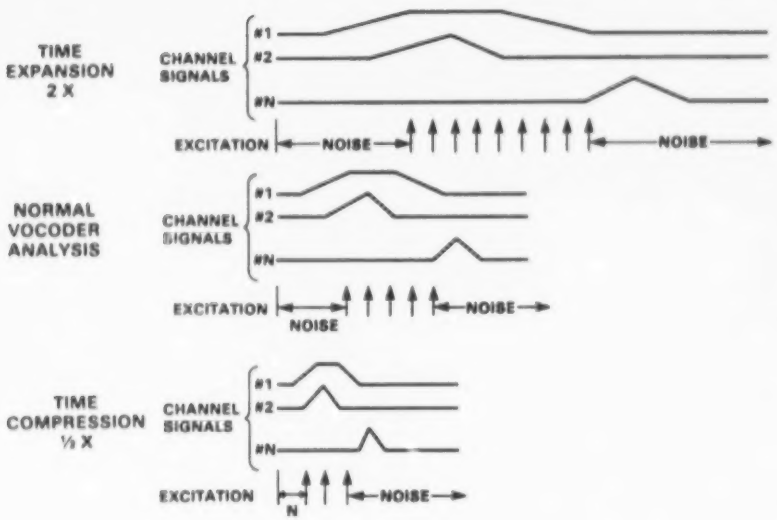


Figure 7. Channel vocoder time expansion and compression.

Instead, a filter based upon transforming the vocal tract spectrum is used to operate on the energy source. This type of analysis-synthesis vocoder has been used to produce acceptable synthetic speech and can thus be used in the same ways as a channel vocoder to produce transformed speech outputs for various communications and research applications.

Linear Predictive Coding Vocoder

Currently, the vocoder which is used most often for speech transmission is the linear predictive coding (LPC) vocoder (see Appendix 2). With this vocoder, the speech signal is generated from the same mechanism and with the same source determination as in the channel and homomorphic vocoders. But with the LPC vocoder, the vocal tract measurement is made by fitting the speech waveform in some interval (usually a frame about 20 msec wide) to the output waveform of a linear filter (Makhoul, 1975).

Fitting a linear filter to the speech waveform can be an efficient process from a computational point of view. As a result, the LPC vocoder is most popular for encoding and transmitting speech at the standard data rate of 2400 bits/sec--the rate used, for example, on a telephone-line modem. One drawback of the LPC vocoder is that, because of the way in which the vocal tract filter is described, the frequency spectrum is not explicitly present for modification. This means that

only a small set of transformations can be implemented or that additional calculations are needed to compute the vocal tract spectrum parameters. Instead of presenting the frequency spectrum explicitly, this vocoder describes the inverse filter impulse response which is used to compute the vocal tract spectrum. This filter description can be used to manipulate certain parameters of the signal directly. For example, it can be utilized to "sharpen" formant frequencies (i.e., to reduce their effective bandwidths) or to reverse the phase of a waveform. If it is necessary to modify only the source signal (as is done in a pitch-perturbation experiment), then the fact that the spectrum is not explicitly present causes no problems. The LPC vocoder may represent the simplest way to implement an analysis for generating modified speech signals. Hence, its application in speech research is potentially great.

Sinusoidal Transform Coding Vocoder

Unlike the vocoders described above, a sinusoidal transform coding (STC) vocoder does not treat the sound source as a periodic pulse train (for voiced sounds) or as a noise signal (for voiceless sounds). Rather, it characterizes the sound source as a sum of sinusoidal waves which vary with respect to one another (Quatieri and McAulay, 1985). In analysis, the peaks of a short-term spectrum measurement are used in specified time-frame intervals to characterize the speech signal sinusoids (see Appendix 3). This "peak picking" occurs in both voiced and voiceless signals. Thus, this coding technique does not separate the sound source into "buzz" (voicing) and "hiss" (voicelessness), one of the more difficult measurement decisions to implement. At all times, the sound source is represented as a sum of sine waves. Provided that a sufficient set of peaks is selected, the set of parameters measured at every frame time characterizes the speech signal completely.

One problem which must be solved in the application of this technique involves matching the measurements made from one frame to the next. It is necessary to match these measurements so that continuous parameter tracks will be generated, i.e., so that the synthesized speech will sound continuous, as natural speech does. A solution to this problem has been developed. It involves an algorithm for a frame-to-frame matching procedure in which the frequencies in frame k are lined up as closely as possible with the frequencies in frame k+1. If there is a peak in frame k that is not matched with a peak in frame k+1, the peak in frame k is allowed to "die." (That is, a zero-amplitude peak of the same frequency is assigned to frame k+1.) Likewise, if there is a peak in frame k+1 that has no counterpart in frame k, its "birth" is assumed. That is, it will be assumed that a zero-amplitude peak of the same frequency started in frame k.[3]

As mentioned in the discussion of the channel vocoder above, it is possible to utilize time expansion and compression with the STC vocoder. However, with this vocoder, time expansion and compression are somewhat more accurate than with the channel vocoder. With such a procedure, it is possible to use speech features themselves to control the time warping so that specific segments of the signal are automatically perturbed or left intact. For example, time expansion (resulting in a slowed speech signal) could occur everywhere in a speech signal except during the production of plosive sounds, such as the stop consonants /t/ and /k/. If these sounds <u>are</u> included in the time expansion, the result is speech which sounds as if the speaker is having difficulty articulating (see discussion of time expansion in the section on channel vocoders above). However, if the time expansion does not include these sounds, the speech sounds normal but deliberate. This kind of

selective transformation of the signal can also be effected by hand-editing a displayed waveform of an utterance. This method would probably be employed in generating stimuli for a perception experiment, since it provides greater control of the transformed utterance than automatic feature-driven time expansion does.

One obvious application of the STC coding procedure in speech research involves utilizing time expansion and compression to determine listeners' sensitivity to rate changes in different segments, such as vowels versus consonants, and in different parts of utterances, such as sentence-initial and sentence-final position. One advantage of using vocoded stimuli here is that fully elaborated sentential and contextualized stimuli could be generated quite easily, enabling one to assess the effect of rate changes in relatively naturalistic utterances.

CONCLUSION

In the present article, it has not been possible to discuss all types of vocoders in use.[4] However, we have presented brief descriptions of four of these vocoders--four which have already been used or could be used in psychophysical and speech-perception research. For, although vocoders were originally developed for the coding and transmission of speech signals, the techniques employed for deriving and transforming these signals can easily be used to produce stimuli for probing listeners' perception. The low data-rate vocoders described above seem especially well suited to applications in language research. Some of the speech-processing techniques we have described are more complex than others, but all have been simulated and implemented in current digital computing facilities--and all are potentially accessible to the interested researcher.

ACKNOWLEDGMENTS

This work was sponsored by the Department of the Air Force. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

THE AUTHORS

Joseph Tierney is a staff member at Lincoln Laborabory, Massachusetts Institute of Technology. His major interests are in speech coding, signal processing, and auditory perceptual mechanisms.

Molly Mack is Assistant Professor in the Division of English as an International Language at the University of Illinois, Urbana. She has worked as a speech researcher and consultant for Lincoln Laboratory, Massachusetts Institute of Technology. Her primary interests are in speech perception and production and in the psycholinguistic and neurolinguistic aspects of bilingualism.

NOTES

[1]Hertz (Hz) is a unit of frequency equal to one cycle per second. For example, if an object vibrates 256 times per second (as does the string on a piano to produce a note of middle C) its frequency is 256 Hz. This value is the fundamental frequency (F0) of the sound. In addition to the F0, a complex sound

such as a piano note (or the human voice) has frequency components or harmonics, each of which has a frequency which is n times that of the F0. Thus, middle C has harmonics at 512 Hz, 768 Hz, 1024 Hz, etc. Further, any complex periodic waveform can be analyzed as the sum of a set of harmonically related sinusoidal (sine) waves. A sine wave appears on an oscilloscope as a simple smooth wave, such as that produced by a tuning fork. The analysis of a complex waveform as the sum of harmonically related sine waves is a fundamental principle of the mathematical procedure called Fourier analysis.

While the fundamental frequency may be considered an acoustic property of the sound source, formant frequencies may be considered acoustic properties of the filter. Formant frequencies (or resonances) are those frequencies at which maximum energy passes through a filter. For example, the first three formant frequencies for [a] produced by an adult male vocal tract are, on the average, 730 Hz, 1090 Hz, and 2440 Hz (Peterson & Barney, 1952).

[2]These and the following figures appear in the appendices. These figures consist of block diagrams of particular vocoders and they are designed to demonstrate (e.g., to a speech-systems computer programmer) how the vocoder algorithms are implemented.

[3]The frame-to-frame frequency matching technique does not, of itself, solve the entire problem of providing a continuous synthesized speech signal. The phase relations must also be interpolated smoothly. Discussion of the procedure for doing so is beyond the scope of this article.

[4]Other types of vocoders being utilized at the MIT Lincoln Laboratory and elsewhere include mid data-rate systems such as the sub-band vocoder and the voice-excited vocoder. In addition, several new techniques for manipulating speech have been developed (Seneff, 1982; Portnoff, 1981). Although the resulting systems produce speech of excellent quality, they are costly in terms of computational complexity compared to the systems described in this article.

REFERENCES

Collier, R. (1975). Perceptual and linguistic tolerance in intonation. International Review of Applied Linguistics, 13, 293-308.

Flanagan, J.L., Schroeder, M.R., Atal, B.S., Crochiere, R.E., Jayant, N.S., & Tribolet, J.M. (1979). Speech coding. IEEE Transactions on Communications, COM-27, 710-737.

Gold, B., Lynch, J., & Tierney, J. 1983. Vocoded speech through fading channels. Proceedings of ICASSP, 83 101-103.

Luce, P.A., Feustel, J.C. & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. Human Factors, 25, 17-32.

Mack, M. In press. The intelligibility of natural and vocoded semantically anomalous sentences: a comparative analysis of English monolinguals and German-English bilinguals. M.I.T. Lincoln Laboratory Technical Report.

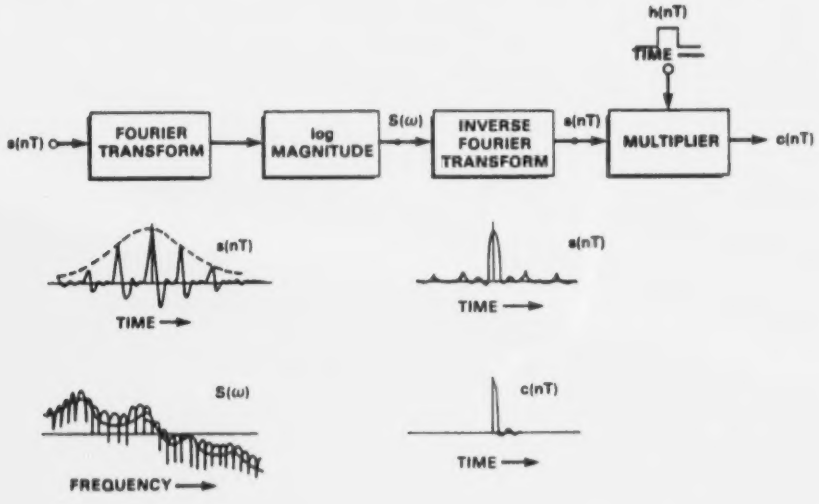
Mack, M. & Gold, B. (1984). The discrimination of pitch in pulse trains and speech. M.I.T. Lincoln Laboratory Technical Report, 680, DTIC No. AD-142996/8.

Mack, M. & Gold, B. (1985). The intelligibility of non-vocoded and vocoded semantically anomalous sentences. M.I.T. Lincoln Laboratory Technical Report, 703, DTIC No. AD-A160401/6.

ct of linguistic content upon the
Mack, M. & Gold, B. (1986). The effe muli. Journal of Phonetics, 14, 333-
discrimination of pitch in monotone sti
337. tutorial review. Proceedings of the
Makhoul, J. (1975). Linear prediction: A
IEEE, 63, 561-580. lysis of perceptual confusions among
Miller, G.A. and Nicely, P.A. (1954). An anal e Acoustical Society of America, 27,
some English consonants. Journal of th
338-352. ysis system based on homomorphic
Oppenheim, A.V. (1969). A speech anal ciety of America, 45, 459-462.
filtering. Journal of the Acoustical So trol methods used in a study of the
Peterson, G.E. & Barney, H.L. (1952). Con ciety of America, 24, 175-184.
vowels. Journal of the Acoustical So nation. Dordrecht: Foris Pub.
de Pijper, J.R. (1983). Modelling British Into tual evaluation of MITalk: The MIT
Pisoni, D.B. & Hunnicutt, S. (1980). Percep E International Conference Record on
unrestricted text-to-speech system. IEE
Acoustics, 572-575. 1985). Perception of synthetic speech
Pisoni, D.B., Nusbaum, H.C. & Greene, B. (IEEE, 73, 1665-1676.
generated by rule. Proceedings of the cation of speech based on short-time
Portnoff, M.R. (1981). Time-scale modific ASSP-29, 3, 374-390.
Fourier analysis. IEEE Transactions, Speech transformation based on a
Quatieri, T.F. & McAulay, R.J. (1985). of ICASSP, 85, 489-492.
sinusoidal representation. Proceedings modify excitation and/or spectrum of
Seneff, S. (1982). System to independently itch extraction. IEEE Transactions,
speech waveform without explicit p
ASSP-30, 4, 566-578. r pitch distance, particularly in speech.
t'Hart, J. (1981). Differential sensitivity for merica, 69, 811-821.
Journal of the Acoustical Society of A J.A., & Aho, E. (1964). Vocoder with
Tierney, J., Gold, B., Sferrino, V., Dumanian, e Acoustical Society of America, 36,
digital pitch extractor. Journal of th
1901-1905.

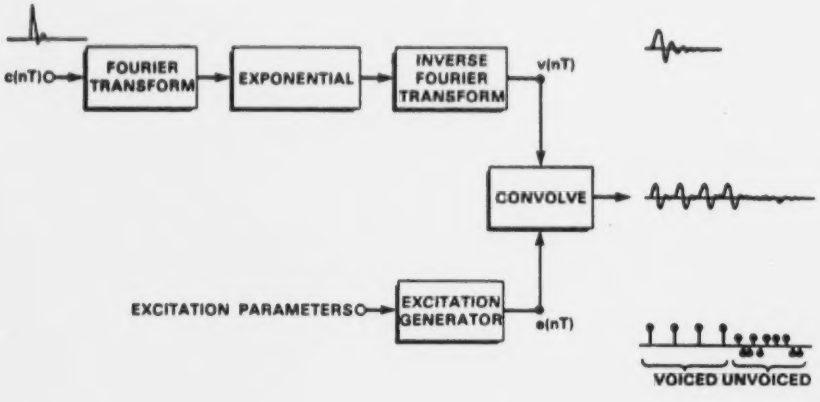
APPENDIX 1a

Homomorphic vocoder: Speech analysis (spectrum separation).



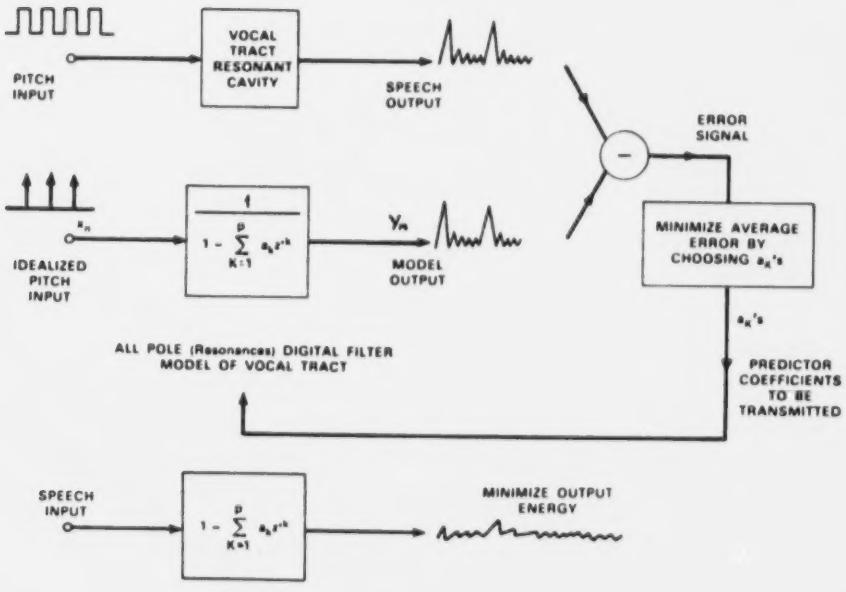
APPENDIX 1b

Homomorphic vocoder: Speech synthesis.



APPENDIX 2

Linear predictive coding (LPC) vocoder: Filter determination.



APPENDIX 3

Sinusoidal tranform coding (STC) vocoder: Analysis and synthesis.